人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-068-04

OWSM v3.1を活用した 既知・未知混合条件における話者ダイアライゼーション

Speaker Diarization in Mixed Known-Unknown Conditions utilizing OWSM v3.1

阿坂 脩平[1*]　Muhammad Shakeel[2]　前田 力[2]　Benjamin Yen[1]
芦澤 剛[1]　住田 直亮[2]　中臺 一博[1]

[1] 東京科学大学
[1] Institute of Science Tokyo
[2] (株) ホンダ・リサーチ・インスティチュート・ジャパン
[2] Honda Research Institute Japan, Co. Ltd.

Abstract: 本稿では，実運用を想定し，既知話者と未知話者が混在する条件における話者ダイアライゼーションの性能向上を目的とする．従来手法では部分的な既知話者情報を活用できないため，本研究では EEND-VC に基づき，以下の 2 点を提案する：1) セグメンテーション性能を底上げする OWSM-Encoder の導入，2) 既知情報を反映する既知シード誘導クラスタリングの導入．実会議コーパスを用いた評価により，OWSM-Encoder が従来モデルと比較して競争力のある結果を示すこと，さらに既知シード誘導クラスタリングの導入によって最大 4.17 ポイントの DER の改善を確認した．

1　はじめに

話者ダイアライゼーションとは,「いつ」「誰が」発話したかを複数話者の会話音声から推定する技術である．高性能な話者ダイアライゼーションを実現することで，その結果を音声認識の補助情報として利用し，高品質な文字起こしを可能にできる．応用例としては，企業等における自動会議録作成が挙げられる．一般的な会議を想定すると，内部参加者については事前に登録された音声特徴を手掛かりとして話者ダイアライゼーションを行える一方，外部参加者の音声特徴は未知であることが多い．話者ダイアライゼーション手法として，ターゲット話者検出に基づく手法 [1, 2] は話者特徴を活用可能だが，全話者の特徴を必要とする．また，近年主流となっている EEND-VC 系列の手法 [3, 4] は事前の話者情報がないことを前提としており，既知情報を活用できない．したがって，既知話者と未知話者が混在する状況で，既知話者の情報を活用できる話者ダイアライゼーション手法の検討が求められる．

本稿では，EEND-VC アーキテクチャを基盤として，部分的な既知情報を活用可能な話者ダイアライゼーションシステムを構築する．具体的には，セグメンテーション部に音声認識タスク向けに事前学習された OWSM [5] のエンコーダを導入し，音響情報に加えて言語的な情報も利用することで，発話区間検知の性能向上を図る．さらに，グローバルクラスタリング部では既知話者の埋め込みを用いた既知シード誘導クラスタリングを導入し，部分的に得られる既知話者情報を活用してクラスタリング性能を向上させる．

評価実験には AMI [6]，AliMeeting [7]，AISHELL-4 [8] の 3 つの遠距離単一チャネル音声コーパスを用いた．その結果，OWSM-Encoder の導入により DER の改善が確認され，他のモデルと比較して競争力のある結果が得られた．また，既知シード誘導クラスタリングにより部分的な既知話者情報を活用することで，最大 4.17 ポイントの DER の改善を確認した．

2　関連研究

話者ダイアライゼーションの従来手法を，既知話者手法と未知話者手法の二つに大別し，それぞれが既知・未知混合条件に有効に対処できない点を指摘する．本稿では，登場話者全員の音声特徴を登録音声として事前に用意し，それを活用して話者ダイアライゼーションを行う手法を既知話者手法とする．一方で，話者数や話者特徴量に関する事前情報を用いず，複数人対話音声のみから発話区間検知を行う手法を未知話者手法と定義する．

*連絡先：東京科学大学
152-8552 東京都目黒区大岡山 2-12-1
E-mail: asaka@ra.sc.e.titech.ac.jp

2.1 既知話者手法

既知話者手法としては，Ding らの Personal VAD [1] や，Medennikov らの Target-Speaker VAD（TS-VAD）[2] が挙げられる．Personal VAD は，事前に取得した音声特徴量を基に，特定話者の発話をフレーム単位で判定する手法である．この考え方は，特定話者を対象とした音声認識 [9] にも応用されている．しかし，一度に一人の話者しか扱えないという制約がある．一方 TS-VAD は，4 人分の音声特徴量を用いて同時に発話区間検知を行い，重複発話にも頑健なダイアライゼーションを実現した．ただし，推論には登場話者全員の音声特徴量が必要である．さらに， TS-VAD 自体は話者特徴推定を含む未知話者手法の研究を目的としているため，既知話者手法としての具体的評価は行われていない．これは，登録音声を利用できる複数話者の実環境コーパスが不足しているためと考えられる．以上より，従来の既知話者手法は未知話者混在条件に対応できず，実環境コーパスを用いた検証も十分ではない．

2.2 未知話者手法

未知話者手法としては，End-to-End Neural Diarization (EEND) [10] が深層学習を活用した手法として有力視された．EEND は同時発話に対応可能な出力形式を備える一方，長時間音声や話者数変動への対応には課題があった．その後，EEND とクラスタリングベースの手法を統合した EEND-VC [3, 4] が木下らによって提案された．この手法では，まずセグメンテーション部において短い音声チャンクごとに EEND を適用し，複数人の重複を含む発話区間を検知する．次に，グローバルクラスタリング部において，その推定発話区間から話者埋め込みを取得する．その後，凝縮型クラスタリングによって推定話者ラベルを付与することで，長時間音声への対応と可変話者数の処理を可能にしている．このアーキテクチャは pyannote フレームワーク [11] を通じて広く利用されており，近年では自己教師あり学習を用いた手法の検討 [12, 13]，音源分離との統合 [14]，さらにクラスタリング手法の改善 [15] などが進められている．しかし，あくまで未知話者手法であるため，本稿が対象とする既知・未知混合条件において，既知話者情報を有効に活用することはできない．

3 提案手法

既知・未知混合条件において部分的な情報を有効に活用できる話者ダイアライゼーションを実現するため，本稿では前節で述べた有力な未知話者手法である EEND-VC を基盤として，二つの提案手法を導入した話者ダイアライゼーションシステムを構築する．提案手法は以下の二点である．第一に，セグメンテーション性能向上のために OWSM v3.1 [5] のエンコーダ部をセグメンテーションモデルへ導入し，言語的な内容に基づく発話区間推定を可能にする．第二に，部分的に得られる既知話者情報を活用するため，既知シード誘導クラスタリングを導入する．本稿では，DiariZen [1] が提供する EEND-VC アーキテクチャを基に，図 1 に示す提案アーキテクチャを構築した．

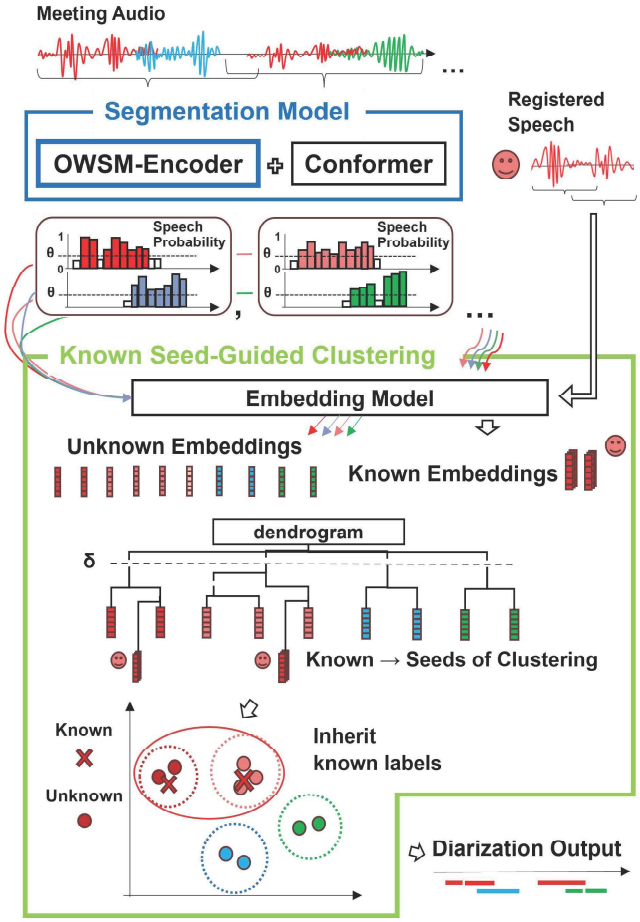

図 1: Proposed Speaker Diarization System

3.1 OWSM-Encoder の導入

既知話者情報の活用効果はセグメンテーション品質に大きく左右されるため，本研究ではまずセグメンテーション性能の向上を目的として OWSM-Encoder を導入する．OWSM（Open Whisper Style Speech Model）[16] は，Whisper [17] と同様の学習をオープンソースツールと公開データのみで再現することを目的としたモデルである．大規模音声データを用いて音声認識およ

[1] https://github.com/BUTSpeechFIT/DiariZen

び翻訳を学習していることから,「汎用音声基盤モデル」としての応用が期待され，実際に幅広い音声タスクで利用されている [18]．参考としたセグメンテーションモデルで用いられている WavLM [12] は，自己教師あり学習を通じて音響的な文脈情報を獲得している一方，言語的・意味的な文脈を明示的に学習しているわけではない．そこで，音声–言語タスクに特化した OWSM に置き換えることで，発話内容に基づく言語的手がかりを活用した発話区間検知が可能になると考えられる．加えて，WavLM と異なり OWSM は事前学習コードが公開されているため，再現性やカスタマイズ性が高い点も利点である．

本稿では E-branchformer [19] を導入することで性能向上が確認された OWSM v3.1 [5] を採用する．導入にあたって注意すべき点として，WavLM は音声波形入力であるのに対し，OWSM のエンコーダ部はメルスペクトログラムを入力とする．そのため，OWSM の入力層で用いられている STFT を併用し，特徴量抽出を担う E-branchformer Encoder とまとめて OWSM-Encoder として導入する．最終的に，セグメンテーションモデルでは OWSM-Encoder による特徴量抽出を行い，その最終層出力を Conformer [20] に入力することで，チャンク単位の推定話者発話区間検知を行う．ただし，音声認識・翻訳向けに学習されたモデルをそのままセグメンテーションに用いることは難しいため，本研究では LibriMix [21] を用い，ESPnet [22] により提供されている話者ダイアライゼーションタスクでファインチューニングを行った．

3.2 既知シード誘導クラスタリング

部分的に得られる既知話者情報を活用して話者ダイアライゼーション性能を向上させるため，既知シード誘導クラスタリングを導入する．従来の EEND-VC における凝縮型クラスタリングは以下の手順で行われる：1) 特徴量抽出器を用いて各チャンクの推定話者の特徴ベクトルを全区間から取得し正規化する，2) ユークリッド距離に基づき樹形図を作成する，3) 距離閾値によりクラスタを生成する，4) 小規模クラスタを最近傍のクラスタに統合する．本手法では，このクラスタリング過程に既知話者特徴ベクトルを組み込む．具体的には，各既知話者の音声のうち指定秒数を抽出し，1) と同じ特徴量抽出器に入力して正規化済みの埋め込みベクトル（既知話者特徴ベクトル）を生成する．既知話者一人あたり複数の特徴ベクトルが得られるが，これらを複製して，同一の特徴量セットを作成する．次に，通常の特徴ベクトルに加えて，既知話者特徴ベクトルも含めた形で樹形図を作成する．この操作により，複製した既知話者特徴ベクトルがクラスタ初期段階でまとまり，クラスタリングのシードとして機能する．さらに，複製倍率が重みとして働き，クラスタの重心を既知話者特徴ベクトル側へ引き寄せる効果が生じ，既知話者情報を基軸としたクラスタリングが可能となる．クラスタ生成後は，既知話者特徴ベクトルを含むクラスタのラベルをその話者に変更し，誤った分割が生じていても同一話者クラスタとして修正可能とする．もし複数の既知話者が同一クラスタに含まれていた場合は，含まれる既知話者特徴ベクトル数が最も多い話者をクラスタラベルとして採用する．このように，従来の凝縮型クラスタリングの枠組みを大きく変更することなく既知話者特徴ベクトルを導入することで，それによるエラーを防ぎつつ，部分的な既知話者情報を活用できる．

4 評価実験

評価実験は二つの観点から行った．実験 1 では OWSM-Encoder 導入による性能向上を評価し，実験 2 では既知シード誘導クラスタリングによる性能向上を評価する．

4.1 データセット

評価には DiariZen の設定と同様に，AMI (AMI-SDM) [6]，AliMeeting [7]，AISHELL-4 [8] の遠距離単一チャンネル音声を用いた．AISHELL-4 のみ，各チャンネル音声を平均した 1 チャンネル音声を使用し，検証データはトレーニングデータの 10% を抽出して作成した．学習・検証データおよびデフォルトの評価データの詳細を表 1 に示す．また，既存の評価データはそのままでは既知・未知混合条件の評価ができないため，構造を再構成したデータを表 2 の通りに用意した．図 2 に示すように，各音声ファイルの冒頭から重複のない発話区間を切り出し，その話者の登録音声として話者ラベルと紐付けて保存した．各ファイルにつき，登場話者 3 名分について，累計 20 秒ずつの登録音声を作成し，その時点で音声を切り分け，残りを評価データとした．AliMeeting については話者数が 3 名未満のファイルを除外し，AISHELL-4 については音声切り出しによって，特定話者の発話が評価データから消失したファイルを除外した．これらの切り出し・除外処理による評価データの縮小率を表 2 に示す．学習と検証には三つのデータセットを併用し，評価は各データセットごとに行う．

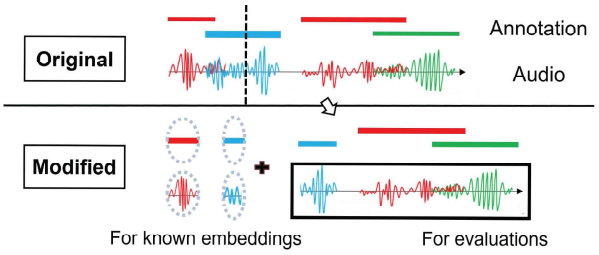

図 2: Method for Creating Evaluation Data

表 1: Dataset

	Train			Dev			Test		
Dataset	files	spk	hrs	files	spk	hrs	files	spk	hrs
AMI	134	3-5	80	18	4	10	16	3-4	9
AliMeeting	209	2-4	111	8	2-4	4	20	2-4	11
AISHELL-4	173	3-7	97	18	3-7	10	20	5-7	13
whole	516	2-7	288	44	2-7	24	56	2-7	33

4.2 学習・推論設定

OWSM-Encoder のファインチューニングには ESPnet [2] を用い，ダイアライゼーションモデルには EEND-EDA [23] を使用した．入力 STFT の設定は FFT 512，窓サイズ 400，フレームシフト 160 とした．学習率は 4e-4，warmup step は 100,000，最適化は AdamW [24] を用い，バッチサイズは 20 とした．データセットは LibriMix [21] を用い，Libri2Mix を 30 epoch，Libri3Mix を 10 epoch 学習し，best モデルを採用した．

セグメンテーションモデルの学習には DiariZen [3] を使用した．比較対象として，ファインチューニングした OWSM v3.1 medium のほか WavLM-base+，WavLM-large を用い，いずれも最終層出力を線形層に通して Conformer に入力した．学習時には事前学習済みモデルも更新対象とし，チャンク 8 秒，ストライド 3 秒，損失関数は Powerset loss [25]，最大話者数は 4 とした．学習率は WavLM ／ OWSM 部分を 1e-5，その他を 1e-3 とし，実効バッチサイズ 64，最適化は AdamW を使用した．最大 100 epoch 学習し，10 epoch 改善がない場合は早期終了した．AutoClip [26] により，勾配ノルムの 90 パーセンタイルに基づきクリッピング閾値を設定した．学習は TSUBAME4.0 node-q（NVIDIA H100 SXM5 94GB）で実施した．

推論時には，事前話者音声を扱えるよう改変した pyannote pipeline [11] を使用し，セグメンテーションモデルは損失が小さい 5 エポックの平均を用いた．推論設定は，チャンク 8 秒，ストライド 0.8 秒，話者数 1〜20，人クラスタリング閾値 0.7，最小クラスタサイズ 30 とした．特徴量抽出には Wespeaker [27, 28] の

[2] https://github.com/espnet/espnet
[3] https://github.com/BUTSpeechFIT/DiariZen

表 2: Modified Test Dataset

	Modified Test					
Dataset	files	known spk	known sec	spk	test hrs	retained rate
AMI	16	3	20	3-4	6.9	76%
AliMeeting	12	3	20	3-4	5.8	53%
AISHELL-4	19	3	20	5-7	9.7	76%
whole	56	3	20	3-7	22.4	69%

ResNet34-LM [4] を使用した．

既知話者特徴ベクトルは，一人当たり 20 秒の登録音声を用い，以下の 3 パラメータで切り出した音声片を特徴量抽出器に通して取得した：

- **Size**: 切り出す音声片の長さ
- **Step**: 音声片を切り出す間隔
- **Replication Factor**: 音声片の複製倍率

本稿では Size = 5.0s，Step = 0.8s を採用した．Size は，チャンク 8 秒内で話者の一貫した発話が概ね 2〜5 秒であること，および既存モデルが 2〜3 秒で学習されていること [29, 30] に基づき，より安定した表現を得るために決定した．Step は，埋め込みの整合性を保つためセグメンテーションのストライドと揃えた．Replication Factor は適切な指標がないため，クラスタリングへの寄与を実験的に探索した．

4.3 評価指標

話者ダイアライゼーションの評価には Diarization Error Rate (DER) を用い，次式のように定義する：

$$\mathrm{DER} = \frac{\text{false alarm}+\text{mis-detection}+\text{confusion}}{\text{total}} \quad (1)$$

ここで，**false alarm** は非発話を誤って発話とした区間，**mis-detection** は発話を誤って非発話とした区間，**confusion** は発話区間における話者誤識別の区間，**total** は全話者の正解発話区間の合計時間である．

本稿では，話者交代時の誤差許容（collar）を設けず，最も厳しい条件で DER を算出する [31]．DER の算出には pyannote-metric [32] を用いる．

4.4 実験 1：OWSM-Encoder の性能比較

OWSM-Encoder の導入によるセグメンテーション性能向上の寄与を評価する．比較手法として WavLM-base+ と WavLM-large を用い，最先端手法との比較のため，未知話者のみの条件であるデフォルトの評価データを使用した．

[4] https://huggingface.co/pyannote/wespeaker-voxceleb-resnet34-LM

4.5　結果 1：OWSM-Encoder の性能比較

比較手法の DER を表 3 に示す．表 3 より，AMI と AliMeeting では WavLM-large には及ばないものの，WavLM-base+ と比較すると競争力のある結果が得られた．AMI において性能向上が得られなかった理由としては，ノイズが大きく環境音が多様であるため，言語的手がかりが十分に作用しなかった可能性が考えられる．一方で AISHELL-4 では，WavLM-base+ より 0.42 ポイント，WavLM-large より 0.31 ポイント高い性能を示した．この結果から，OWSM-Encoder による言語的・意味的文脈の反映は，対象とする音声によっては有効に働くことが示された．

4.6　実験 2：提案クラスタリングの性能比較

既知シード誘導クラスタリングの導入による性能向上を評価する．未知・既知混合条件を扱うため，4.1 節で述べた，構造を変更したデータセットを用いる．まず，ベースラインとして WavLM-base+ を用い，Replication Factor を実験的に探索する．パラメータの候補は [5,10,20,50,100,200,400,750] とし，それぞれ既知話者が 1 人の場合と 3 人の場合について DER を算出した．次に，その結果に基づき，既知話者特徴ベクトルの抽出パラメータを確定し，WavLM-base+，WavLM-large，OWSM-Encoder を用いた場合の，既知話者人数による DER の比較を行った．既知話者の組み合わせが複数ある場合は全通りを実行し，平均値を算出した．また，2 節の通り，既知・未知混合条件で比較可能な従来手法が存在しないため，既知話者情報を利用しない場合をベースラインとする．

4.7　結果 2：提案クラスタリングの性能比較

Replication Factor を変化させたときの DER を表 4 に示す．これより，既知話者人数に依らず，Modified AMI では Replication Factor = 5〜10 が適しており，他の二つのコーパスでは Replication Factor = 100〜200 が適していることが読み取れる．したがって，既知話者特徴ベクトル抽出パラメータを (Size, Step, Replication Factor) = (5, 0.8, 10) および (5, 0.8, 200) の 2 種類に決定した．

これらのパラメータを用いた場合の既知話者人数ごとの DER を表 5 に示す．結果より，全体として既知話者情報を適用した方が，適用しない場合より良い性能を示す傾向が確認できる．しかし，既知話者人数が多いほど性能が向上するとは一概には言えず，既知話者特徴ベクトル間で分布が重なりエラーを生むケースがあると考えられる．個別のデータセットを見ると，AMI

(a) Without known information

(b) 3 Known Speakers Applied

図 3: Clusters of R8002_M8003_MS803 in AliMeeting

では二つのパラメータに共通して，WavLM-large の 1 既知話者適用時が最高性能であり，未適用時と比較して 0.26 ポイント改善した．AliMeeting では，Replication Factor = 200 かつ WavLM-base+ の 3 既知話者適用時が最高性能で，未適用時より 4.17 ポイント改善した．ここでは，Replication Factor が大きい，つまりは既知話者特徴ベクトルの重みが大きい方が，AliMeeting 全体として既知話者適用時の改善幅が大きい傾向が示された．これは，AliMeeting が中国語による激しいインタラクションを特徴とする複雑なコーパスであり，既知話者情報という明確な基準が，話者識別において有効に作用したためと考えられる．また，AISHELL-4 では，Replication Factor = 200 かつ WavLM-base+ の 2 既知話者適用時が最高性能で，未適用時より 0.12 ポイント改善した．ただし，既知話者人数によっては改善と同程度のエラーも発生している．これらの結果より，コーパスの特徴に応じて，既知シード誘導クラスタリングの効果が大きく左右されることがわかった．

さらに，既知シード誘導クラスタリングの効果を視覚的に確認するため，UMAP [35] による可視化を行った．図 3 には，性能向上が大きかった AliMeeting における最終的なクラスタ分布を示す．これより，既知話者を適用することで，従来は混同していた話者クラスタが，既知話者特徴ベクトルを手掛かりとして分離できていることが確認され，提案手法の有効性が裏付けられた．

表 3: Performance Comparison of OWSM-Encoder

Model	AMI				AliMeeting				AISHELL-4			
	DER	FA	MD	CN	DER	FA	MD	CN	DER	FA	MD	CN
WavLM-base+	15.54	3.69	8.26	3.59	18.26	2.52	8.41	7.33	10.67	4.10	3.57	3.00
WavLM-large	**14.55**	3.76	7.40	3.39	**15.11**	2.46	8.00	4.65	10.56	4.81	2.69	3.06
OWSM-Encoder	16.06	3.53	8.99	3.54	16.95	2.52	8.44	5.99	**10.25**	4.59	2.58	3.08
SOTA	14.0 [33]				12.5 [34]				9.8 [33]			

表 4: Experimental Exploration of Replication Factor

	Replication Factor								
num_known = 1	0	5	10	20	50	100	200	400	750
Modified AMI	16.58	**16.56**	16.57	16.61	16.61	16.61	16.62	16.62	16.62
Modified AliMeeting	24.75	24.80	23.24	23.25	23.23	23.04	**22.44**	22.61	22.61
Modified AISHELL-4	12.34	12.33	12.38	12.31	**12.24**	**12.24**	12.25	**12.24**	12.31
	Replication Factor								
num_known = 3	0	5	10	20	50	100	200	400	750
Modified AMI	16.58	**16.49**	**16.49**	16.64	16.65	16.65	16.65	16.65	16.66
Modified AliMeeting	24.75	22.43	22.44	22.42	22.41	20.58	20.58	20.59	**20.57**
Modified AISHELL-4	12.34	12.31	12.26	**12.25**	12.26	**12.25**	12.26	12.26	12.47

5 おわりに

本稿では，既知話者と未知話者が混在する状況で，部分的な情報を活用できる話者ダイアライゼーションを実現することを目標とした．そのために，EEND-VC を基に，セグメンテーション性能向上のための言語的文脈を扱う OWSM-Encoder の導入と，部分的既知話者情報を活用出来る既知シード誘導クラスタリングの導入を行ったアーキテクチャを提案した．実会議コーパスを用いて，提案手法の評価を行ったところ，OWSM-Encoder は従来モデルと競争力のある結果を示し，既知シード誘導クラスタリングの導入によって，DER の改善を確認した．今後は，コーパスに性能を左右されない手法の提案と，セグメンテーション部への既知情報の活用を目指す．

参考文献

[1] Ding, S. et al.: Personal VAD: Speaker-Conditioned Voice Activity Detection, *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pp. 433–439 (2020)

[2] Medennikov, I. et al.: Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario, *INTERSPEECH*, pp. 274–278 (2020)

[3] Kinoshita, K. et al.: Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, *ICASSP*, pp. 7198–7202 (2021)

[4] Kinoshita, K. et al.: Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech, *arXiv preprint arXiv:2105.09040* (2021)

[5] Peng, Y. et al.: OWSM v3.1: Better and faster open whisper-style speech models based on e-branchformer, *arXiv preprint arXiv:2401.16658* (2024)

[6] Carletta, J. et al.: The AMI meeting corpus: A pre-announcement, *International Workshop on Machine Learning for Multimodal Interaction*, pp. 28–39 (2006)

[7] Yu, F. et al.: M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge, *ICASSP*, pp. 6167–6171 (2022)

[8] Fu, Y. et al.: Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario, *arXiv preprint arXiv:2104.03603* (2021)

[9] Maeda, C. *et al.*: Joint Target-Speaker ASR and Activity Detection, *Proc. INTERSPEECH*, pp. 1683–1687 (2025)

[10] Fujita, Y. et al.: End-to-end neural speaker diarization with self-attention, *IEEE ASRU*, pp. 296–303 (2019)

[11] Bredin, H.: pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe, *INTERSPEECH*, pp. 1983–1987 (2023)

[12] Chen, S. et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing, *IEEE JSTSP*, Vol. 16, No. 6, pp. 1505–1518 (2022)

[13] Han, J. et al.: Leveraging self-supervised learning for speaker diarization, *ICASSP*, pp. 1–5 (2025)

表 5: Performance Comparison of Proposal Clustering Methods

With Params: (Size, Step, Replication Factor) = (5, 0.8, 10)

Model	Modified AMI				Modified AliMeeting				Modified AISHELL-4			
num_known	0	1	2	3	0	1	2	3	0	1	2	3
WavLM-base+	16.58	16.57	16.52	16.49	24.75	23.24	22.45	22.44	12.34	12.38	12.35	**12.26**
WavLM-large	16.17	**15.91**	15.93	15.96	**21.73**	22.27	22.08	22.67	12.35	12.62	12.48	12.28
OWSM-Encoder	17.26	17.35	17.37	17.38	22.67	22.89	22.71	23.25	12.28	12.45	12.41	12.29

With Params: (Size, Step, Replication Factor) = (5, 0.8, 200)

Model	Modified AMI				Modified AliMeeting				Modified AISHELL-4			
num_known	0	1	2	3	0	1	2	3	0	1	2	3
WavLM-base+	16.58	16.62	16.62	16.65	24.75	22.44	21.03	**20.58**	12.34	12.25	**12.22**	12.26
WavLM-large	16.17	**15.91**	15.92	15.94	21.73	21.43	20.60	20.74	12.35	12.52	12.41	12.37
OWSM-Encoder	17.26	17.36	17.37	17.37	22.67	22.47	21.92	22.08	12.28	12.30	12.26	12.29

[14] Kalda, J. et al.: PixIT: Joint training of speaker diarization and speech separation from real-world multi-speaker recordings, *Odyssey*, pp. 115–122 (2024)

[15] Delcroix, M. et al.: Multi-stream extension of variational bayesian HMM clustering (MS-VBx) for combined end-to-end and vector clustering-based diarization, *arXiv preprint arXiv:2305.13580* (2023)

[16] Peng, Y. et al.: Reproducing whisper-style training using an open-source toolkit and publicly available data, *IEEE ASRU Workshop*, pp. 1–8 (2023)

[17] Radford, A. et al.: Robust speech recognition via large-scale weak supervision, *PMLR International Conference on Machine Learning*, pp. 28492–28518 (2023)

[18] Shakeel, M. *et al.*: Unifying Diarization, Separation, and ASR with Multi-Speaker Encoder, *Proc. ASRU* (2025)

[19] Kim, K. et al.: E-branchformer: Branchformer with enhanced merging for speech recognition, *IEEE SLT Workshop*, pp. 84–91 (2023)

[20] Gulati, A. et al.: Conformer: Convolution-augmented transformer for speech recognition, *arXiv preprint arXiv:2005.08100* (2020)

[21] Cosentino, J. et al.: LibriMix: An Open-Source Dataset for Generalizable Speech Separation, *arXiv preprint arXiv:2005.11262* (2020)

[22] Watanabe, S. et al.: ESPnet: End-to-End Speech Processing Toolkit, *arXiv preprint arXiv:1804.00015* (2018)

[23] Horiguchi, S. et al.: End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, *arXiv preprint arXiv:2005.09921* (2020)

[24] Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, *arXiv preprint arXiv:1711.05101* (2017)

[25] Plaquet, A. et al.: Powerset multi-class cross entropy loss for neural speaker diarization, *INTERSPEECH*, pp. 3222–3226 (2023)

[26] Seetharaman, P. et al.: Autoclip: Adaptive gradient clipping for source separation networks, *IEEE MLSP Workshop*, pp. 1–6 (2020)

[27] Wang, H. et al.: Wespeaker: A research and production oriented speaker embedding learning toolkit, *ICASSP*, pp. 1–5 (2023)

[28] Wang, S. et al.: Advancing speaker embedding learning: Wespeaker toolkit for research and production, *Speech Communication*, Vol. 162, pp. 103104 (2024)

[29] Desplanques, B. *et al.*: ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification, *arXiv preprint arXiv:2005.07143* (2020)

[30] Snyder, D. *et al.*: X-Vectors: Robust DNN Embeddings for Speaker Recognition, *Proc. ICASSP*, pp. 5329–5333 (2018)

[31] Landini, F. et al.: Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks, *Computer Speech & Language*, Vol. 71, p. 101254 (2022)

[32] Bredin, H.: pyannote. metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems, *INTERSPEECH*, pp. 3587–3591 (2017)

[33] Han, J. et al.: Efficient and Generalizable Speaker Diarization via Structured Pruning of Self-Supervised Models, *arXiv preprint arXiv:2506.18623* (2025)

[34] Plaquet, A. *et al.*: Dissecting the Segmentation Model of End-to-End Diarization with Vector Clustering, *arXiv preprint arXiv:2506.11605* (2025)

[35] McInnes, L. et al.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction *arXiv preprint arXiv:1802.03426* (2018)